MATE-T580 Practical Data Science using R
Assignment 1
Comparing Healthcare Outcomes of Developed Countries

You often hear in the news that the USA spends more on healthcare than developed countries with comparable or better healthcare system. For this assignment, your task is to do some preliminary analysis regarding this claim. Since measuring healthcare expenditure and healthcare outcomes can be quite tricky, for this exercise, we'll agree on the following two metrics:

- Healthcare expenditure is measured as ***share of GDP*** spent on healthcare.
- Healthcare outcome is measured as ***average life expectancy at birth***.

You are provided with two datasets (in the form of .csv files) obtained from the Organisation for Economic Cooperation and Development (OECD): the first contains information on healthcare expenditure, and the other on healthcare outcomes. The two datasets contain much more information than needed for this assignment. Your task is to clean, combine, and manipulate the datasets as such to produce a single R data frame with the following specifications:

- Each row of the output data frame corresponds to a country.
- The data frame has four columns:
  - Country: The country
  - Spending: Total amount spent on healthcare as % share of GDP
  - Life: Average life expectancy at birth in years (a single value reflecting both male and female populations)
  - Outcome: Ratio of Life expectancy to Spending
- The data frame is sorted according to Outcome by descending order (i.e. the top country is the one that gets the most relative to what it spends and likely the bottom country is the one that gets the least relative to what it spends).

Here are some tips to help you get started:

- After loading the datasets in R, and before doing any manipulations to the datasets, start by inspecting them using the *str* or *glimpse* functions.
- Next, reduce the number of rows in each dataset according to the metrics of interest for this analysis. Remember that we don't care about all measures of healthcare expenditure and all measures of healthcare outcome, only the two defined above.
- Limit the analysis to year 2015.

- For the healthcare outcomes table, life expectancy for males and females are listed separately, you need to combine these into a single measure. The *dplyr* functions *group_by* and *summarize* can by handy here.
- Join the two datasets by Country and retain only those countries that have records in both datasets. This operation can be performed using the *inner_join* function.
- Create a new variable Outcome, representing the ratio between life expectancy and share of GDP spent on healthcare.
- Sort the data frame according to Outcome (from highest to lowest).
- Inspect the result by looking at the top 5 and bottom 5 rows. Do they make sense to you?
- (Optional) You may do some basic plotting to gain more insights into the data.

Appendix

The csv files can be downloaded from:

https://drive.google.com/uc?export=download&id=1yKjNfEtjmNYJBXLUfCjtzkaIrPElzZT7

https://drive.google.com/uc?export=download&id=1VSoN8CVRsQKluNAS8lNYPxJAeeEo7bV-